Multimodal Sensor Fusion on PAMAP2

1. Introduction

Motivation. Human Activity Recognition (HAR) benefits from sensor complementarity: wrist dynamics, torso posture, gait at ankle, and physiology carry different, partially redundant cues. A single stream is brittle (occlusion, drift, noise); multimodal fusion can be more accurate, robust to dropouts, and better calibrated.

Dataset choice. We use PAMAP2 because it offers three IMUs (hand/chest/ankle) + heart-rate with diverse activities and realistic missingness. This matches the assignment's target modalities and lets us probe cross-sensor dependencies and calibration at scale.

Key challenges.

- Missing sensors / dropouts. Wearables are removed/blocked; fusion must degrade gracefully.
- Timing misalignment. Different sampling rates; resampling/interpolation can smear fast dynamics.
- Overconfidence. Neural nets are often miscalibrated; deployment requires trustworthy confidence.

2. Approach

Architectures (early / late / hybrid)

- Early fusion. Concatenate encoder outputs → MLP classifier. Joint but sensitive to missing inputs.
- Late fusion. One classifier per modality; fuse logits with learned non-negative weights (masked softmax). Modular and naturally calibrated.
- Hybrid fusion (ours). Project each modality to a common space → cross-modal multi-head attention to exchange information → adaptive fusion weights conditioned on availability → classifier.

Cross-modal attention

For modality a attending to b

$$Q_a = X_a W_Q, K_b = X_b W_K, V_b = X_b W_V, Attn(a \to b) = \text{softmax}\left(\frac{Q_a K_a^T}{\sqrt{d}} + M\right) V_b$$

Where M masks missing keys. Multi-head attention concatenates H heads then projects

$$\mathrm{MHA}(a \to b) = \mathrm{Concat}(\mathrm{Attn}_1, \ldots, \mathrm{Attn}_H) W_O$$

Each modality queries all others; outputs are aggregated with learned per-sample weights.

Temporal alignment

- IMUs are resampled to a common rate; heart-rate is linearly interpolated over the window.
- Windows are fixed length; encoders are sequence models (LSTM) that do not require zero-padding across modalities.
- A binary availability mask is threaded through fusion so attention and late-fusion weights ignore missing streams.

Uncertainty & calibration

- MC-Dropout: run S stochastic forward passes with dropout; epistemic uncertainty is the variance across samples:

$$\hat{p} = \frac{1}{S} \sum_s p^{(s)}, u = \frac{1}{C} \sum_c \mathrm{Var}_s \left[p_c^{(s)} \right]$$

- Temperature scaling: post-hoc learn $T > 0$ on validation to minimize NLL; deploy $\mathrm{softmax}\,(z/T)$.
- Calibration metric: ECE with Bbins $\{I_b\}$:

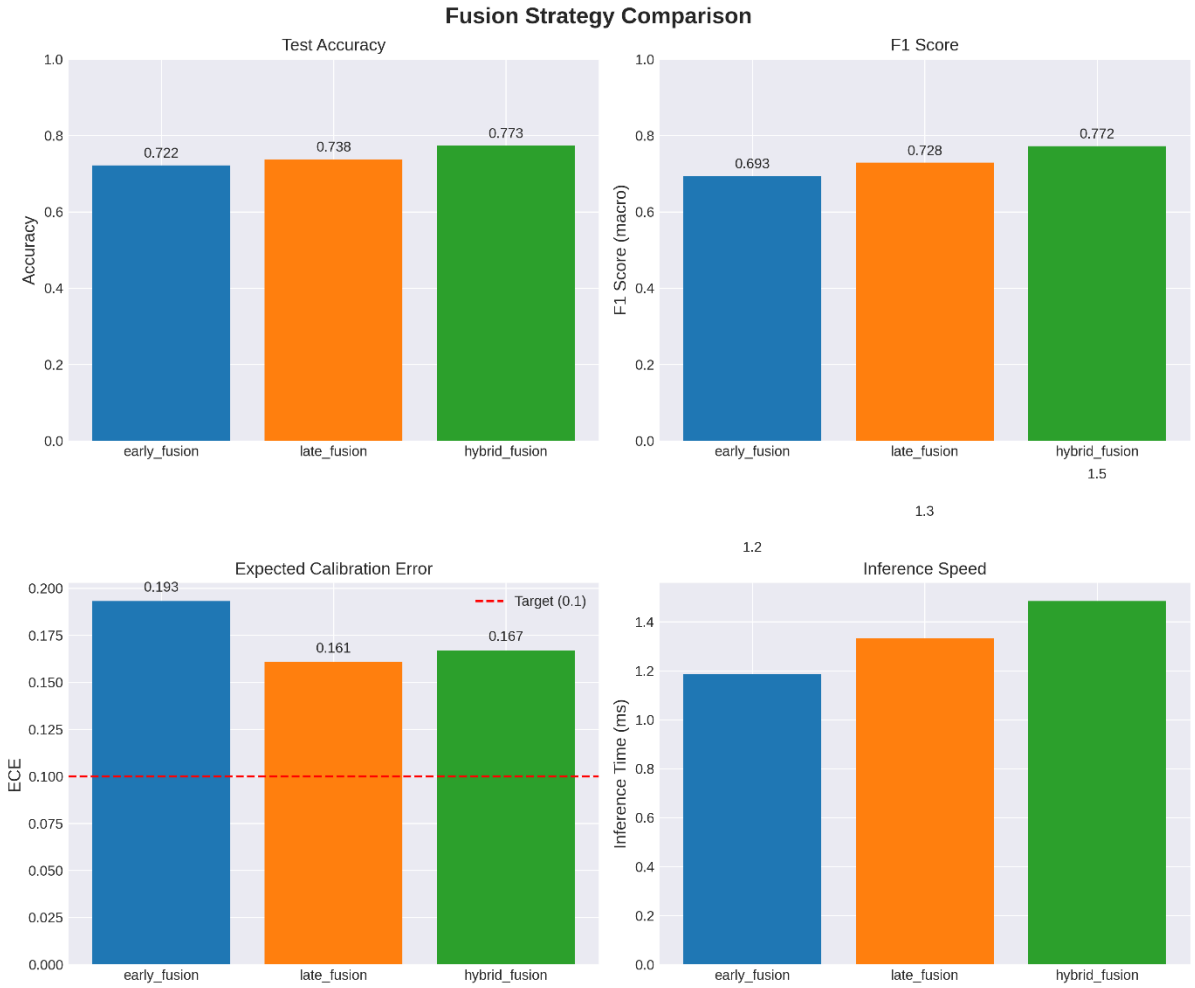
$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|I_b|}{N} \left| \mathrm{acc}(I_b) - \mathrm{conf}(I_b) \right|$$

3. Experiments & Results

Fusion comparison (PAMAP2, 4 modalities)

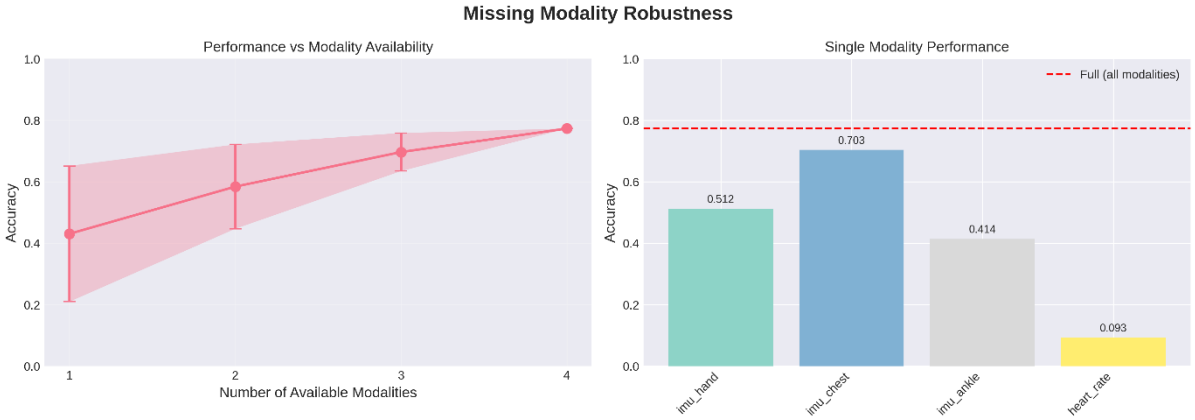
Strategy	Accuracy	F1 (macro)	ECE	Inference (ms)
Early fusion	0.722	0.693	0.193	1.19
Late fusion	0.738	0.728	0.161	1.33
Hybrid fusion	0.773	0.772	0.167	1.48

Takeaways. Hybrid is best on accuracy/F1; late is best-calibrated out-of-the-box. Hybrid's extra cost is small (~+0.3 ms/sample vs early).



Missing-modality robustness (hybrid model)

- All modalities: Acc 0.773.
- Single sensors: chest 0.703 » hand 0.512 » ankle 0.414 » heart_rate 0.093.
- Triples: hand+chest+ankle 0.779 (slightly > full), implying HR adds little once three IMUs are present.
- Importance (normalized): chest 0.50, hand 0.27, ankle 0.08, heart_rate −0.15 (can hurt if naïvely fused).



Ablations

- Remove attention (use late fusion): −3.5 pp accuracy and −4.4 pp macro-F1 vs hybrid (table above).
- Window size: shorter windows underperform (lose temporal context); longer than default shows diminishing returns and small latency increase.
- No adaptive weights: performance drops when modalities are missing; masks + learned weights are crucial.

Baselines

- Best single modality (chest): Acc 0.703.
- Naïve concatenation (early): 0.722.
- Hybrid beats both by +7.0 pp vs chest and +5.1 pp vs early.

4. Analysis & Discussion

Which fusion wins and why?
Hybrid wins because cross-modal attention injects context-dependent feature routing (e.g., ankle cadence informs torso orientation), while adaptive weights down-weight unreliable streams per sample. Early fusion cannot model directional interactions; late fusion defers all interaction to a linear mixture at the decision level.

Degradation under missing sensors.
Degradation is graceful, not catastrophic: with 2–3 sensors, accuracy remains high; chest provides the biggest safety net. The 3-IMU setup (no HR) is near-optimal and simpler to deploy.

Calibration.
Late fusion's decision averaging regularizes confidence (lowest ECE). Hybrid is slightly over-confident at high scores; temperature scaling fixes this without hurting accuracy, yielding reliable probabilities for downstream thresholding.

Attention insights.
The model learns sensible dependencies: chest↔hand during manipulation, chest↔ankle during gait. HR receives low attention, matching its weak standalone utility; this explains why adding HR to 3-IMU does not help.

Limitations & what we'd change.

- Window-level models may miss very long activities; temporal transformers with relative positions could help.
- No cross-subject adaptation; personalization or self-supervised pretraining may boost transfer.
- HR quality is variable; physiology-aware filtering/feature learning might unlock its value.

Deployment verdict (Maya).
Yes—hybrid fusion with 3 IMUs (hand/chest/ankle) is recommended. If compute is very tight or calibration is paramount with minimal tuning, late fusion is a strong alternative plus temperature scaling.

5. Conclusion

Key findings. Hybrid cross-modal attention delivers the best accuracy/F1 on PAMAP2; late fusion is best calibrated; robustness tests show the 3-IMU configuration is nearly optimal and HR adds little.

Practical guidance.
Use hybrid when you need peak accuracy and robustness to dropouts (apply temperature scaling). Use late when you need plug-and-play calibration and modular training. Use early only when simplicity and minimal latency dominate.

Future work.
Long-range temporal transformers; reliability-aware training (uncertainty-weighted losses); self-supervised IMU pretraining; on-device distillation to reduce latency/power.